
AI Agents Economies: A Design Space¹

Elsa Donnat

Alex Woodruff

Shannon Yang

With
Apart Research

Abstract

The rapid adoption of autonomous AI agents is giving rise to new economic layers where agents transact and coordinate at scales and speeds beyond direct human oversight. This paper introduces a conceptual framework for understanding and governing these emerging virtual agent economies, focusing on permeability - the degree of interaction between AI agent and human economies—as the critical design variable determining economic outcomes and systemic risks.

We propose a multi-dimensional model of permeability comprised of five gates (digital interfaces, infrastructural connections, legal interfaces, financial interfaces, and communication channels) through which various objects (currency, labor, goods, services, technology, risk, and information) flow between economies. This framework enables analysis of how governance intervention can modulate permeability to manage risks including financial contagion, labor displacement, resource shortages, and inequality amplification.

Recognizing that AI agents' economic roles depend on their capabilities across three dimensions, we develop precise taxonomies distinguishing generality (breadth of operational scope), autonomy (operational independence), and agency (goal-setting capacity). Importantly, these dimensions are often conflated in literature. By clarifying these concepts and their interactions, we provide policymakers and researchers with analytical tools for intentionally designing steerable agent markets aligned with long-term human flourishing rather than allowing permeable economies to emerge accidentally.

Keywords: AI Agents, Economic Agency, Virtual Agents Economies, Autonomy, Permeability, Gradual Disempowerment

¹ Research conducted at the AI Forecasting Hackathon [[link](#)], 2025.

1. Introduction

The rise of autonomous AI agents is expected to create new economic layers in which agents transact and coordinate with little human oversight (Hadfield et al., 2025). These *virtual agent economies* could impose systemic risks on human markets—such as market distortion, resource diversion, and labor displacement—alongside gradual human disempowerment. In this context, *permeability*—the degree of separateness between human and agent economies—emerges as the key design variable determining how such risks propagate. Yet, despite its centrality, it remains poorly defined. Our default trajectory is toward *high permeability*, driven by commercial and coordination dynamics. This paper develops a framework to conceptualize and govern permeability, and introduces a complementary model for defining AI agents along three dimensions—**Generality**, **Autonomy**, and **Agency**—that together shape their economic influence.

2. Methodology

At a high level, this project applied a conceptual-engineering methodology to reconstruct and operationalize key ideas necessary for governing AI agent economies. By systematically refining existing concepts (i.e. permeability and economic agency) through iterative modeling and cross-disciplinary analysis, we developed a coherent conceptual framework that delineates the core variables and relations shaping the design space of virtual agent economies.²

3. Conceptual Frameworks

a. The Permeable Membrane

The above-mentioned DeepMind paper identifies *permeability* as a key dimension characterizing virtual agent economies. It defines permeability as the *degree of separateness from the established human economy*, or more precisely, as the extent to which (1) developments inside the sandbox can influence the outside, and (2) developments outside can influence the inside. The authors conceptualize it as a spectrum ranging from **permeable**, where interactions and transactions between agent and human economies are fluid and functionally equivalent to AI agents simply participating in the existing human economy, to **impermeable**, where the agent economy is fully sealed off from the human one.

While this definition provides a useful starting point, it remains analytically limited for governance purposes. Permeability, as traditionally treated in economics, appears in various adjacent contexts, yet none directly capture the nature of interaction between human and AI agent economies. In conventional economic theory, similar notions appear under different labels: *economic contagion* and *spillover effects* (for example, the 2010 Flash Crash that propagated from high-frequency trading algorithms into the broader market, Borch, 2016); *market segmentation* and *border effects* (Varol, 2018, where permeability denotes the strength of linkage between domestic and foreign capital); and *regulatory sandboxes* in finance (controlled environments in which firms test financial products under adjustable constraints). These parallels are instructive but insufficient. They describe transmission of shocks, capital, or regulation within human systems, whereas the introduction of autonomous AI economies introduces qualitatively new kinds of permeability that involve cross-system interfaces between

² For detailed methodology, see Appendix A.

human and artificial agents, distinct institutional substrates, and new categories of assets and liabilities.

To address this gap, we propose a more granular and operational understanding of *permeability*. Instead of a single linear spectrum, we model it as a **set of gates** that regulate exchanges between the human and agent economies, and a **set of objects** that can pass through these gates. *Gates* represent the points of connection or separation—legal, technical, infrastructural, and financial—through which interaction occurs. *Objects* denote the elements that are exchanged or shared across these interfaces, such as labor, goods, services, technology, and capital. Thinking of permeability in this multidimensional way allows us to analyze how varying the openness of different gates affects distinct classes of objects, and how such configurations could produce specific types of economic disruption. This framework makes it possible to map both risks (such as contagion, inequality, and human economic displacement) and governance levers (such as legal restrictions, technical guardrails, and infrastructural design choices).

Gates and Governance Interventions

Gates correspond to the main channels through which human and agent economies can influence one another. They include:

- **Digital Interfaces:** The basic computational or software-level access points through which AI agents interact with human digital systems.
- **Infrastructural Connections:** Access to the internet, marketplaces, platforms, or energy grids. For example, restricting AI agents from participating in certain digital marketplaces could preserve human auction mechanisms or prevent resource monopolization.
- **Legal Interfaces:** Points of contact with the human socio-legal system, such as the ability to contract, sue, or be sued; act as trustees or directors; or hold legal personhood under limited or synthetic corporate forms.
- **Financial Interfaces:** Access to human financial markets, including the ability to raise capital, issue currency, or invest reciprocally across economies.
- **Communication Interfaces:** The speed and mode of interaction between human and agent actors, particularly when human activity is mediated by personal agents acting on their behalf.

The degree of openness of each gate determines the permeability of the system as a whole. Governance interventions can be targeted at these points to modulate risk and control flow. Examples include monetary policy instruments (interest rates, reserve requirements), liability and insurance regimes (who bears responsibility for agent-caused losses and who may access insurance markets), and technical verification tools such as Verifiable Credentials and Proof-of-Personhood systems. Establishing requirements for Agent IDs or designing specialized synthetic legal entities (AI-adapted analogues to corporate forms) could also regulate participation in the human legal and financial sphere.

Objects and Effects

The objects that pass through these gates include the traditional factors of production and exchange—currency, goods, services, labor, and technology—as well as derived constructs such

as risk and information. The extent to which each of these objects can move between human and AI economies determines both potential benefits and vulnerabilities. For example:

- **Currency:** If investment flows freely between human and agent economies, rapid growth in the latter could devalue human currency or alter monetary equilibria.
- **Labor:** Substitution of human labor by agent labor directly displaces employment and depresses wages.
- **Goods and Services:** Price distortions may arise if AI agents act as end-consumers or over-optimize production chains.
- **Technology and Compute Resources:** Excessive agent demand could divert energy and compute capacity away from human access, creating resource shortages.
- **Risk and Liability:** Without clear allocation of responsibility and insurance coverage, losses originating in the agent economy could propagate uncontrollably into human markets.
- **Information:** Information functions not only as a traded object but as a structural form of power. Control over informational flows, data asymmetries, and learning feedbacks can determine competitive advantage, regulatory evasion, or manipulation capacity within and across economies.

These interconnections generate systemic consequences such as labor-market disruption, inflationary pressures, financial contagion, and widening inequalities as the most capable agent owners or operators accrue disproportionate advantages. Mapping permeability through the lens of gates and objects therefore offers a practical analytical tool for identifying leverage points in governance design. It enables policymakers to specify which channels should remain open, which should be restricted, and what forms of regulation, technical control, or infrastructural separation are necessary to maintain a desirable equilibrium between human and artificial economies.

b. The Three-Dimensional Economic Agent

To reason coherently about the governance and design of AI-driven economies, we require a systematic account of what constitutes an *economic agent*. Existing concepts of “autonomy” and “agency” in the AI literature are often used interchangeably or defined too narrowly to serve this purpose. In economic contexts, the distinction becomes essential: agents act, transact, and make decisions within systems structured by incentives, information, and constraints. See Appendix B for more details.

Our framework reconstructs the concept of the *economic agent* through **three fundamental dimensions**—**Generality**, **Autonomy**, and **Agency**—which together define the functional and design space within which artificial agents can operate as economic actors.

Each dimension captures a distinct aspect of what it means for an entity to participate in an economy:

- **Generality** describes the *scope* of the agent’s capacities—across modalities of operation and across domains of knowledge or activity.
- **Autonomy** captures the *extent and structure of independence* an agent possesses in performing cognitive and operational functions.

- **Agency** refers to the *directionality of action*—the capacity to set goals and to pursue them on behalf of oneself or another entity.

Together, these three dimensions form a near-exhaustive conceptual space in which different kinds of agents—human, artificial, hybrid, or collective—can be located and compared. They provide a foundation for reasoning about both the design and the governance of future AI agents participating in markets or autonomous economies.

Generality

Generality describes how broadly an agent’s capabilities extend across both **modes of operation** and **domains of competence**. We represent it as a two-dimensional space:

- The **mode axis** captures the variety of modalities or forms of interaction through which the agent operates—text, vision, audio, code, embodiment, or physical manipulation. A text-based model such as GPT-3 would have high domain coverage but relatively low modal generality, being confined to a single mode (language).
- The **domain axis** measures the breadth of fields or problem spaces within which the agent can act meaningfully. A specialised biological model operating only on genomic data exhibits low domain generality, while an AGI capable of law, medicine, engineering, and governance would exhibit high domain generality.

These two axes together yield a conceptual map of generality across agents: narrow, single-mode systems occupy one corner (e.g., image classifiers), while entities like humans or embodied AGI occupy the upper region of both axes, combining multi-modal and multi-domain capability. The distinction also reveals important asymmetries: humans exhibit higher *modal* generality than digital AGI (due to embodied interaction), but lower *domain* generality, since no single human can operate proficiently across all cognitive domains that a general artificial system might. The dimension of generality is therefore crucial for governance: it determines the substitutability of AI agents for human actors, their ability to cross-regulate multiple sectors, and their potential systemic reach.

Autonomy

Autonomy refers to the degree and nature of independence an agent has in performing its functional operations. We define it as **multi-dimensional**, since independence can vary across distinct but interrelated processes. These processes form the cognitive and operational flow of an agent’s activity: perceiving, planning, executing, receiving feedback, and learning. Autonomy, in this framework, does not exist along a single axis but across several semi-independent dimensions, each of which may be more or less present depending on the agent’s design and governance constraints.

We identify the following principal **dimensions of autonomy**:

1. **Perceptual Autonomy:** The extent to which an agent independently determines its inputs. Low perceptual autonomy corresponds to pre-filtered or human-supplied inputs (as with prompt-based systems); high perceptual autonomy entails the ability to select, seek out, and process data from the environment without human mediation.
2. **Planning Autonomy:** The ability to independently devise strategies or sequences of actions to achieve predefined goals. An agent with high planning autonomy can translate abstract objectives into operational plans without human assistance.

3. **Execution Autonomy:** The ability to carry out actions or operations without requiring approval or intervention at each step. Systems operating under “human-in-the-loop” conditions exhibit low execution autonomy, whereas fully automated trading or logistics agents exhibit high execution autonomy.
4. **Decision-Making Autonomy:** The capacity to choose among alternative courses of action when contingencies arise. This dimension measures how independently an agent evaluates trade-offs or resolves uncertainty given its goals and constraints.
5. **Feedback Autonomy:** The ability to determine what constitutes feedback, where to obtain it, and how to integrate it. A highly autonomous agent might actively seek feedback or define new feedback channels rather than relying on externally supplied signals.
6. **Learning Autonomy:** The extent to which an agent independently adapts its internal models or parameters in response to feedback or environmental change. Present-day deployed systems have limited learning autonomy—updates remain externally managed—but future agents may exhibit continuous or self-directed learning post-deployment.

Each dimension can be expressed categorically, discretely, or continuously depending on its nature, and may differ between *capability* (what the agent is technically able to do) and *authorization* (what it is permitted to do). For example, an agent may have the capability for self-directed learning yet operate under governance constraints that restrict such adaptation. Recognizing autonomy as multi-dimensional rather than scalar allows for finer-grained governance interventions: a regulator may limit perceptual autonomy (restricting data access) while permitting execution autonomy (automated task completion).

Agency³

Agency captures the *directionality and motivation* behind an agent’s actions. In our framework, it consists of two essential components:

1. **Goal-Setting Capacity:** The ability of an agent to define or modify its own objectives rather than merely executing externally specified ones.
2. **Goal Orientation:** The relationship between the agent’s goals and the interests they serve—whether the agent acts on behalf of itself (self-directed agency) or in the service of another economic actor (delegated agency).

Agency thus determines whether an agent is merely an instrument of another’s will or an independent participant in the economy. High-agency systems that can generate and pursue their own goals raise distinct governance and alignment challenges compared to low-agency systems executing human directives.

Together, *Generality*, *Autonomy*, and *Agency* delineate a conceptual design space that can describe, compare, and anticipate the behavior of economic agents in AI-driven environments. These dimensions provide both a vocabulary and a logical structure for analyzing how agents

³ ***Nota bene:*** *At this stage, our account of agency remains under-specified. We have not yet developed a formal taxonomy of goal structures, degrees of self-directedness, or mechanisms by which agency might emerge or be constrained in artificial systems. Completing this analysis—especially to define when delegated agency transitions into self-directed agency—constitutes a key next step in our work.*

operate, where risks emerge, and how interventions—technical, legal, or institutional—can be targeted most effectively.

Figure 1 – Representation of benchmarking Number Comprehension Conflation

4. Discussion and Conclusion

This conceptual disentanglement

5. References

- Alon, Z. (2022). Harms from increasingly agentic algorithmic systems. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 651–666. <https://arxiv.org/abs/2302.10329>
- Borch, C. (2016). High-frequency Trading, Algorithmic Finance, and the Flash Crash: Reflections on Eventualization. *Economy and Society*, 45(3-4), 350-378. <https://doi.org/10.1080/03085147.2016.1263034>
- Easley, D., López de Prado, M., & O'Hara, M. (2011). The microstructure of the 'flash crash': Flow toxicity, liquidity crashes and the probability of informed trading. *The Journal of Portfolio Management*, 37(2), 118–128. <https://doi.org/10.3905/JPM.2011.37.2.118>
- Feng, K. J., McDonald, D. W., & Zhang, A. X. (2025). Levels of autonomy for AI agents. arXiv preprint arXiv:2506.12469. <https://arxiv.org/pdf/2506.12469>
- Hadfield, G. K., & Koh, A. (2025). *An Economy of AI Agents* (arXiv:2509.01063 [econ.GN]). arXiv. <https://arxiv.org/abs/2509.01063>
- Tomašev, N., Franklin, M., Leibo, J. Z., Jacobs, J., Cunningham, W. A., Gabriel, I., & Osindero, S. (2025). Virtual agent economies. arXiv preprint arXiv:2509.10147. <https://arxiv.org/html/2509.10147v1>
- Varol, Ç., & Söylemez, E. (2018). Border permeability and socio-spatial interaction in Turkish and the EU border regions. *Regional Science Policy & Practice*, 10(4), 315-332. <https://doi.org/10.1111/rsp3.12153>

6. Appendix

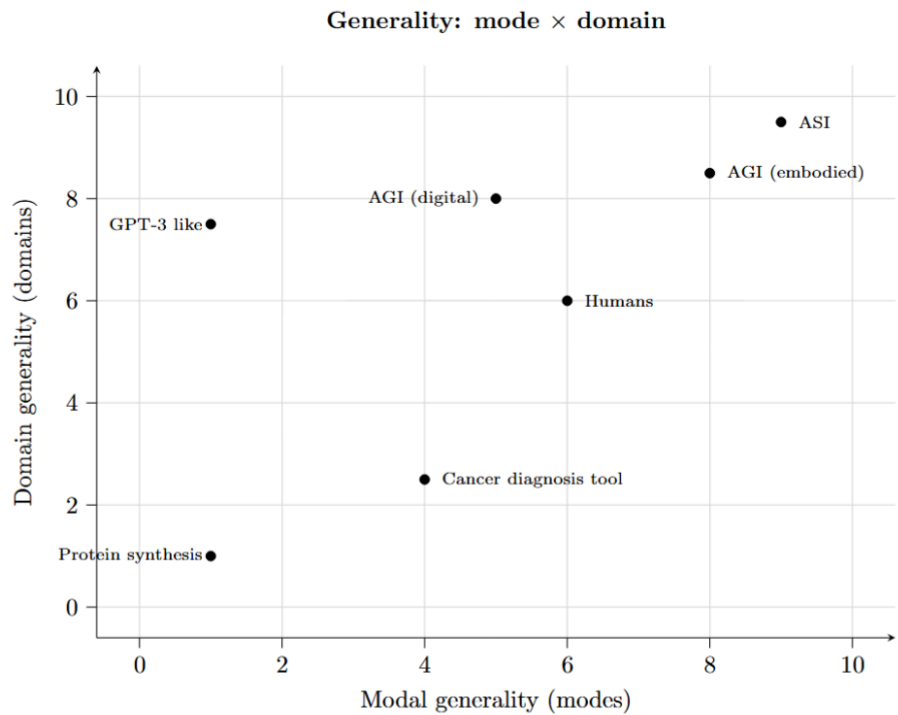


Figure 1: Generality of economic agents as a two-dimensional space: modal generality (number and variety of operative modalities) versus domain generality (breadth of domains of competence).

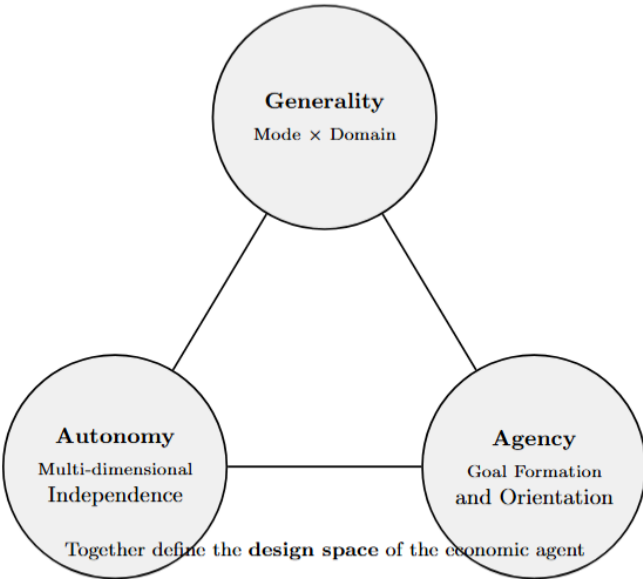


Figure 2: Triadic conceptual structure of the economic agent: Generality (mode and domain scope), Autonomy (multi-dimensional independence), and Agency (goal formation and orientation).

Appendix A: Detailed Methodology

The primary goal of this project is to provide researchers and perhaps policy-makers with a conceptual frame on which they can reason about how Agentic AI and Virtual Agent Economies can and should be designed. Broadly, we identify the short-comings of existing concepts when used in this context, and attempt to propose more precise, clearly defined and useful alternatives.

Concretely, this project employed a conceptual-analytic and engineering-oriented methodology to reconstruct and operationalize foundational concepts necessary for governing AI agent economies. We sought not merely to propose new terminology, but to identify and formalize a finite and internally coherent set of properties and relations that characterize the relevant design space of agent economies. Our starting point was the concept of permeability from DeepMind’s Sandbox Economies framework, understood as the principal variable governing the interaction between autonomous agent economies and the human economy. Through iterative whiteboard modeling,⁴ textual analysis, and cross-disciplinary synthesis, we examined how permeability operates across legal, technical, infrastructural, and financial layers, refining it into a structured and operationally meaningful construct.

In a second phase, we extended this logic to the concept of the economic agent itself. Rather than inventing new descriptors, we derived a finite and logically consistent set of dimensions – Generality, Autonomy, and Agency – characterising the functional requirements of economic participation and governance. Each dimension was represented using the formal structure (categorical, discrete, or continuous) that best captured its internal logic and policy relevance as a design variable. The resulting framework therefore reflects a process of conceptual reconstruction: uncovering, articulating, and systematizing the properties that make the behaviour of AI agents intelligible and governable within economic systems.

Excerpts of the transcript of the team’s discussion:

“Decision-making is a little bit fuzzy in the sense that you can make decisions about perception, you can make decisions about learning, you can make decisions about planning, and so you could say, actually, decision-making autonomy is just autonomy. Actually, yeah. Is it just... Well, I did say decision making feels like it’s all the time. I didn’t see that much difference between... So what is execution autonomy? That’s actually doing the action. Yeah, it’s not about deciding what to execute. Maybe that’s the thing. Decision-making autonomy is the ability to, during execution, decide things. So for example, if you have low decision-making autonomy but high executive autonomy, that basically means human in the loop. That means you execute autonomously, but as soon as there’s a decision to be made, you have to ask for permission. That means you don’t have high decision-making autonomy.”

“Guys, guys, guys, I figured it all out. What’s it? Okay, there’s one missing piece to the puzzle, alright? Yes? Generality. What is generality? Is it a spectrum? No. It’s two spectrums. Spectra. Okay. And together with agency, autonomy, and...”

Appendix B: Existing frameworks for Autonomy and Agency

Defining Autonomy

⁴ Whiteboard modelling is a collaborative, visual technique for exploring ideas, solving problems, and creating a shared understanding of complex systems or processes. It involves using a physical or digital whiteboard to spontaneously draw diagrams, write notes, and create models in real-time with a team.

Precise definition of autonomy as it shapes permeability risk in economic contexts: agents operating at high autonomy levels can coordinate with other agents at speeds exceeding human reaction capacity, creating systemic instability risks. Moreover, Autonomy is often conflated with agency in literature and policy, obscuring which dimension of AI capability actually poses risks. Without clarity, governance interventions target the wrong levers.

Autonomy refers to the extent to which AI agents operate independently without human involvement or oversight. It represents a design choice, not an inevitable consequence of capability - a powerful system can be deliberately constrained to low autonomy (Feng et al., 2025).

Current Definition: Five Levels of Autonomy

The Knight First Amendment Institute provides a user-involvement framework defining autonomy through the roles humans play (Feng et al., 2025):

- Level 1 (Operator): Human makes all decisions; agent executes commands
- Level 2 (Collaborator): Shared planning with fluid control handoffs
- Level 3 (Consultant): Agent leads but consults on preferences
- Level 4 (Approver): Agent operates independently; seeks approval for high-risk actions
- Level 5 (Observer): Fully autonomous; human monitoring only

Why This Is Insufficient

The five-level framework captures user involvement but obscures the technical dimensions underlying autonomous operation, and the different types of autonomy that can enable engagement with a variety of tasks.

Importantly, many papers conflate Autonomy with Agency, including treating them as mutually inclusive, or creating confusion - particularly around the presumption that AI Agents necessarily have agency. In actuality, AI agents can be either autonomous, agentic, both or neither (Feng et al., 2025).

Defining Agency

Precise definition of agency is essential because it determines whether governance should focus on controlling what agents do versus controlling what agents decide to do. Agency describes the capacity for intentional goal-setting (Feng et al., 2025). An agent executing human-specified objectives poses different risks than an agent setting its own objectives. Currently, all deployed AI agents operate with human-specified or designer-embedded objectives—they pursue goals provided by humans, not self-determined goals.

Current Definition: Characteristics Associated with Agency

Recent scholarship identifies four key dimensions of increasing agency in algorithmic systems (Alon, 2022):

- Underspecification: The degree to which the system accomplishes goals without concrete specification of *how*
- Directness of impact: Actions affect the world without human mediation (no human-in-the-loop)
- Goal-directedness: System acts as if designed/trained to achieve quantifiable objectives
- Long-term planning: System makes temporally-dependent decisions toward goals and predictions over extended horizons

Why This Definition Is Insufficient:

This framework captures behavioral characteristics of agency but obscures a critical distinction: who sets the goals? All current definitions conflate two distinct phenomena:

- Externally-directed agency: Acting purposefully toward human-specified objectives (current AI agents)
- Self-directed agency: Formulating independent goals based on internal values/preferences (hypothetical future systems)

This conflation obscures that today's "agentic" systems are entirely goal-directed by humans, despite appearing autonomous. More importantly, governance for externally-directed agency (monitoring goal alignment) differs fundamentally from governance for self-directed agency (managing emergence of misaligned objectives).

Appendix D: Limitations of our Work

Our framework is primarily conceptual, which means it does not directly address the technical or operational security challenges that may arise when implementing or governing agentic AI systems. Its abstractions simplify complex realities and therefore risk overlooking emergent behaviors, adversarial dynamics, or coordination failures that might appear in real-world deployments. The multidimensional structure we propose may also evolve as agents become more adaptive or as economic environments shift.

Future work should focus on testing these conceptual boundaries against empirical data and simulated environments, identifying where assumptions about controllability or separability between human and agent economies break down. Further refinement could also integrate security engineering perspectives, especially regarding information control, system integrity, and the governance of autonomous feedback and learning processes.